Proceedings of International Conference on Software Engineering (ISE'06)
14th – 15th April, 2006, Lahore, Pakistan

A Web Smart Space Framework for Information Mining: A base for Intelligent Search Engines

M. Asif Naeem,
Balochistan University of Information Technology and Management Sciences
Quetta, Pakistan
naeembuitms@hotmail.com

Imran S. Bajwa,
Balochistan University of Information Technology and Management Sciences
Quetta, Pakistan
imransbajwa@yahoo.com

Riaz-Ul-Amin,
Balochistan University of Information Technology and Management Sciences
Quetta, Pakistan
riazulamin@hotmail.com

M. Abbas Choudhary
Balochistan University of Information Technology and Management Sciences
Quetta, Pakistan
abbas@buitms.edu.pk

Abstract

A web smart space is an intelligent environment which has additional capability of searching the information smartly and efficiently. New advancements like dynamic web contents generation has increased the size of web repositories. Among so many modern software analysis requirements, one is to search information from the given repository. But useful information extraction is a troublesome hitch due to the multi-lingual; base of the web data collection. The issue of semantic based information searching has become a standoff due to the inconsistencies and variations in the characteristics of the data. In the accomplished research, a web smart space framework has been proposed which introduces front end processing for a search engine to make the information retrieval process more intelligent and accurate. In orthodox searching anatomies, searching is performed only by using pattern matching technique and consequently a large number of irrelevant results are generated. The projected framework has insightful ability to improve this drawback and returns efficient outcomes. Designed framework gets text input from the user in the form complete question, understands the input and generates the meanings. Search engine searches on the basis of the information provided.

Keywords: Search Engine Preprocessing, Web Smart Space, Natural Language Processing

1. Introduction

Search Engine is software that searches for data based on some criteria. Every Web search engine site uses a search engine that it has either developed itself or has purchased from a third party. Search engines can differ dramatically in the way they find and index the material on the Web, and the way they search the indexes from the user's query. Although a search engine is technically the software and use algorithms to perform a search. For example, Google is a major search site on the Web, but rather than being called the "Google search site," it is commonly known as the "Google search engine."

Smart Spaces needs a methodology to deal with complexity. Our approach is to construct smart spaces from autonomous parts, which we call agents, responsible for their own data and actions and communications with each other [4]. This gives a high intrinsic adaptability and survivability (desirable properties unless we wish to employ a multitude of maintenance engineers for every smart space) and encourages what are called "emergent behaviors" due to the combined actions and interactions of many agents. A smart space [3] is an environment with numerous elements that:

- can sense
- can think
- can act
- can communicate
- can interact with other people

This intelligent environment is also robust, self-managing, and scaleable. Driven by rapid developments in Web searching technology and Web mining, such smart spaces can play an important role to perform useful tasks and to tackle the complex issues on web.

Natural language processing (NLP studies the problems of automated generation and understanding of natural human languages. Natural language generation systems convert information from computer databases into normal-sounding human language [8], and natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate.

2. Related Work

Information gathering on the Internet is a time consuming and somewhat tedious experience. The main method for gathering information on the Internet is the search engine. Literature reviewed indicates that Intelligent Information

Agents can improve the process of gathering information on the Internet [1].

In general, when searching for a piece of information in a Web page, blocks likely to contain desired information are searched first, then more fine-grained blocks recursively until the desired information is found. Instead of recursively searching within a page, hierarchical information searching is a more straightforward process for finding and understanding page information [7]. General search engines always crawl and index everything in a page and decide which words or paragraphs are more important by analyzing term frequency and inverse document frequency [12]. In the general design of information agents, human's knowledge is first called for deciding which blocks are needed for agents to crawl. It is difficult for general search engines and information agents to find important information for users.

Usually, Agents are used for information retrieval from web. But it is difficult for them to recognize the significance of information in a page and they always consider the contents of Web pages as a linear data stream[9], where no such consideration of contextual information exists. The relations among different contexts and links in a web page is a significant element which increases accuracy and decreases the cost of search engines during the extraction and classification of informative regions of a web page. The structure of information within a page to represent the significance and relation of information is called the information hierarchy of a page.

Agents collaborate to gather HTML pages from de the World Wide Web and treat them in order to be able to retrieve those pages from subsequent users' queries. Crawling Agent collaboration is required in order to decide the URLs that should be first retrieved. Subsequent age treatment consists on first filtering the pages so that HTML format is transformed into XML and second indexing them so that information retrieval can be performed online [2]. A search engine operates, in the following order

- Web crawling
- Deep Crawling Depth-first search (DFS)
- Fresh Crawling Breadth-first search (BFS)
- Indexing
- Searching

Web search engines work by storing information about a large number of web pages, which they retrieve from the WWW itself. These pages are retrieved by a web crawler (sometimes also known as a spider) an automated web browser which follows every link it sees [6], exclusions can be made by the use of robots.txt. The contents of each page are then analyzed to determine how it should be indexed (for example, words are extracted from the titles, headings, or special fields called meta tags). Data about web pages is stored in an index database for use in later queries. Some search engines [13], such as Google, store all or part of the source page (referred to as a cache) as well as information about the web pages, whereas some store every word of every page it finds, such as AltaVista. This cached page always holds the actual search text since it is the one that was actually indexed, so it can be very useful when the content of the current page has been updated and the search terms are no longer in it.

3. Description of Problem

The problem specifically addressed in this research is primarily related to a large number of the irrelevant outcomes of a searching query. In conventional techniques, search engines use pattern matching methodologies and the web contents are searched by matching user given words. This surface matching technique generates results in millions and billions and most of the time irrelative and unrelated results are shown. To make search activity more efficient and effective, the major emphasis of research is contents searching bodies as information agents. Mostly research is being done at backend in the form of multi agents. The preprocessing of user's query before a search engine processes it is also significant.

4. Proposed Solution

In this research paper a framework is proposed, known as web smart space that accentuates upon front end. Designed framework gets text input from the user in the form of complete question, understands the input and generates the meanings. Search engine searches on the basis of the information provided after pre-processing of user's query. In the context of this research, objects are automatically identified from a problem domain. User provides the input text in English language according to his requirement. After the lexical analysis of the text, syntax analysis is performed on word level to recognize the word's category [7]. First of all the available lexicons are categorized into nouns, pronouns, prepositions, adverbs, articles, conjunctions, etc. The syntactic analysis of the programs would have to be in a position to isolate subjects, verbs, objects, adverbs, adjectives and various other complements. It is little complex and multipart procedure. "Who is the inventor of Computer?"

For this example, following is the output.

Lexicons	Phase-I	Phase –II
Who	subject	entity
is	helping-Verb	method
inventor	co-subject	entity

the	article	-------
computer	object	entity

Table-1.0 A user's query analysis

This is the final output of lexical assessment phase and all nouns are marked as objects and verbs are marked as methods and all adjective are marked as states of that particular object. In the above example, there are two subjects 'who' and 'inventor', 'is' is method of subject 'who', 'the' is the article and 'computer' is object.

5. Natural Language Processing

The natural languages are irregular and asymmetrical. Usually, natural languages are based on un-formal grammars. There are the geographical, psychological and sociological factors which influence the behaviors of natural languages [8]. There are indeterminate set of words and they also change and vary area to area and time to time. Due to these variations and discrepancies, the natural languages have different flavors as English language has more than half dozen famous flavors all over the world. These flavors have different accents, set of vocabularies and phonological aspects. These warnings and ominous discrepancies and contradictions in natural languages make them a complex task to process them as compared to the formal languages [14].

In the procedure of analyzing and understanding the natural languages, different troubles are generally faced by the researchers. The problems concerned to the greater complexity of the natural language are verb's conjugation, inflexion, lexical amplitude, problem of ambiguity, etc. From this set of problems the problem which still causes more difficulties is problem of ambiguity. Ambiguity could be easily solved at the syntax and semantic level by using a powerful and robust rule-based system.

6. Used Methodology

Conventional natural language processing based systems use rule based systems. Agents are another way to develop speech language based systems [16]. In the research, a rule-based algorithm has been designed and used which has robust ability to read, understand and extract the desired information. First of all, basic elements of the language grammar are extracted [2] as verbs, nouns, adjectives, etc then on the basis of this extracted information further processing is performed. In linguistic terms, verbs often specify actions, and noun phrases the objects that participate in the action [7]. Each noun phrase's then role specifies how the object participates in the action.

A procedure that understands such a sentence must discover the subject because he performs the action of invention [8], the computer is the thematic object because it is the object that is invented. Thus, complete sentence analysis finds information about the subject, co-subject, object, recipient, etc. The identification of such information specifically helps to understand the meanings of the input sentence as given below.

Subject: The representative causes the action to occur as in "Mary Bellis invented the computer" Mary Bellis is representative who performs the task. But in this example a passive vice sentence, the representative also may appear as "The computer is invented by Mary Bellis." In the context of this research paper only active vice sentence have been attempted.'

Co-Subject: If representative is working with any other partner that is called co-representative. Both of them carry out the action together as "Mary Bellis and Douglas Engelbart are working" In this sentence Marry Bellis is subject and Douglas Engelbart is Co-subject.

Recipient: The receiver is the person for whom an action has bee performed: "Alexander Fleming discovered penicillin for humanity" In this sentence humanity is receiver and usually receiver come after 'for' preposition.

Object: The object or thematic object is a thing; the sentence is really all about— typically the object, undergoing a change. Often the thematic object is the same as the syntactic direct object, as "Mary Bellis invented the computer." Here the computer is thematic object.

Position: The position is where an action occurs. Several prepositions are manifesting the position usually a noun phrase as "Windows operating system was designed in Bell Laboratory." Here Bell Laboratory is position and come after 'in', 'on' and 'under' prepositions.

Time: Time specifies when an action occurs. Prepositions such at, before and after introduce noun to depict time as "The first prototype computer mouse was made to use with a graphical user interface (GUI) after 1964." Here 1964 represents time.

Period: Period or duration specifies how long an action takes. Preposition such as since and for indicate duration. "Integrated circuits are being used since 1956." Here the time from 1965 to now is period or duration.

Route: Motion from one point to pother or from source to destination takes place over a route or trajectory. It contrast to the other role possibilities, several prepositions can serve to introduce trajectory noun phrases: "What is air route from Pakistan to China." Here 'from' and 'to' prepositions have been used for route.

7. Proposed System Anatomy

The proposed system, "A Web Smart Space Framework for Information Mining: A base for Intelligent Search Engines " consists of four layers that are Natural Language Processing, Web mining agents, knowledge representation and Web information repositories. Here is the detail description of the framework layers. The first layer of proposed framework is Natural Language Processing (NLP). Following is the architecture of the designed system for a Web Smart Space Framework for Information Mining

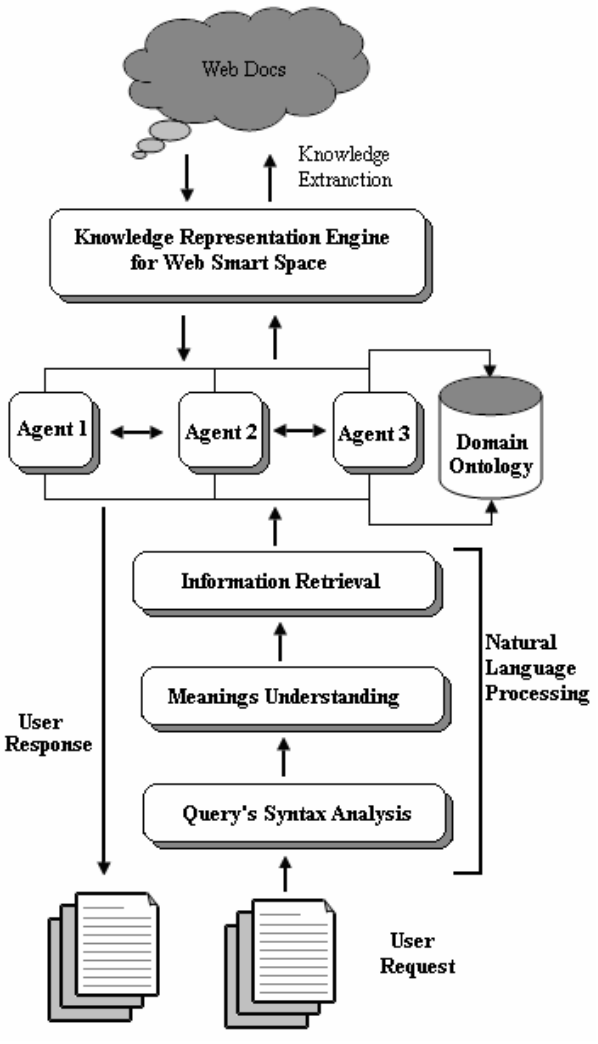

Figure 1.0: A Web Smart Space Framework for Information Mining

This layer further distributed into further three layers that are Query's Syntax Analysis, Meaning Understanding and Information Retrieval. First of all user request is differentiated into list of words. After the lexical analysis of the text, syntax analysis is performed on word level to recognize the word's category. All the available lexicons are categorized into nouns, pronouns, prepositions, adverbs, articles, conjunctions, etc. The syntactic analysis of the programs would have to be in a position to isolate subjects, verbs, objects, adverbs, adjectives and various other complements. After analyzing the available nouns, pronouns, prepositions, adverbs, articles and conjunctions their meaning is to be understood and then finally the required information is retrieved from given request. Second layer consists of distributed agents and domain ontology, the specific data region related to given request. During the process of searching these distributed agents can communicate with each other and can share data related to request. By implementing this technique more refined results can be produced quickly. Third layer of proposed frame work is knowledge representation engine. These knowledge bases actually save and update the web history automatically. First of all search engine search the request from this knowledge-base and in case of failing request is searched from repositories. Fourth layer of proposed system is web information repositories, which are the collection of various physical locations, where web information are stored and later retrieved.

8. Information Agents

Various types of agents are used for various purposes as interface agents, task agents, information agents and middle agents. Interface agents typically interact with the users, receive user input, and display results [4]. The task agents basically help users perform tasks, formulate problem-solving plans and carry out these plans by coordinating and exchanging information with other software agents and on the other hand the middle agents help match agents that request services with agents that provide services [5].

Information agents typically provide intelligent access to a heterogeneous collection of information sources. Various set of functions are performed by the agents. The Communication and Coordination module accepts and interprets messages and requests from other agents [6]. The Planning module takes as input a set of goals and produces a plan that satisfies the goals. The Scheduling module uses the task structure created by the planning module to order the tasks. The Execution module monitors this process and ensures that actions are carried out in accordance with computational and other constraints.

According to diagram each agent has three major parts. First is domain knowledge, which helps it in finding the required result for assigned request. Second and most important feature is goal or target that an agent has to

complete by doing successful searching and third is input and output channel through which one agent can communicate to another agent.

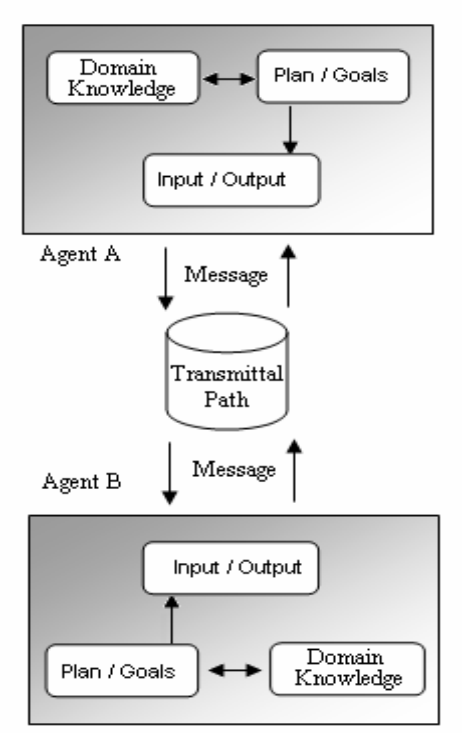

Figure 2.0: Distributed Information Agents communicating with each other

Figure 2.0 represents the internal structure as well as method of communication between two information agents.

9. Conclusion

In this paper, A Web Smart Space Framework for Information Mining: A base for Intelligent Search Engines has been presented. The proposed framework bases on a Natural Language Processing techniques and has a perceptive ability to understand the user requirements and then search and extract the refined results from the given Web repository. A new idea of web smart space has been introduced which basically introduces front end processing for a search engine to make the information retrieval process more intelligent and accurate. As in common searching techniques, searching is performed only by using pattern matching technique and consequently a large number of irrelevant results are generated. The projected framework has insightful ability to improve this drawback and returns efficient outcomes. Designed framework gets text input from the user in the form complete question, understands the input and generates the meanings. This preprocessed information is used to extract only the required information from web.

10. Future Work

The Web Smart Space Framework for information mining using Natural Language Processing may be used to get the more useful information from Web repository. Future struggle in this regard is to implement the proposed framework like Google search engine. The designed algorithm is used to understand the user's information and extract the relative information. Current algorithm only considers the active-vice sentences. Improvement in the algorithm can improve the accuracy ratio of the result and can ultimately influence the accuracy ratio of searching results.

11. References

[1] Francisco J. Martín, Enric Plaza, Juan Antonio Rodriguez-Aguilar, "An infraestructure for Agent-Based Systems: An Interagent Approach". *International Journal of Intelligent Systems (John Wiley & Sons, Inc.),* vol 15, pages 217-240, 2000.

[2] M. López-Sánchez, F. Martín, J. García, X. Canals, X. Drudis, N. Ruiz, and A. Reyes , Agent Communication within a Search Engine Architecture

[3] Sven van der Meer, Brendan Jennings, Keara Barrett, Ray Carroll (2003), Design Principles for Smart Space Management, *1st International Workshop on Managing Ubiquitous Communications and Services (MUCS)*

[4] Naomi Augar. Intelligent information agents: search engines of the future

[5]. Yang, P. Pai, V. Hanovar, L. Miller, (1998) Mobile Intelligent Agents for Document Classification and Retrieval: A Machine Learning Approach, *Proceedings of the European Symposium on Cybernetics and System Research*, Vienna, Austria

[6] H.S.Nwana, M.Wooldridge, Software Agent Technologies, Software Agents and Soft Computing: Towards Enhanced Machine Intelligence, *Lecture Notes in Artificial Intelligence* 1198, pp.59-77, 1997.

[7]. Tang L, Rong A, Yang Z. A review of planning and scheduling systems and methods for integrated steel production, *European Journal of Operational Research 2001*; 133(1):1-20.

[8]. Vaessen N, Van Nerom L, Beghin P. Production optimisation through integrated caster and hot strip mill scheduling. *Proceedings of the 2nd International Conference on Production Planning and Control in the Metals Industry, London, UK;* 1996, p. 21-26.

[9] Ouelhadj D, Cowling PI, Petrovic S. Utility and stability measures for agent-based dynamic scheduling of steel continuous casting. *Proceedings of the IEEE International Conference on Robotics and Automation, Taiwan ; 2003*, p. 175-180.

[10] Ouelhadj D, Cowling PI, Petrovic S. Contract net protocol for cooperative optimisation and dynamic scheduling of steel production. In: Ibraham A, Franke K, Koppen M, editors. *Intelligent Systems Design and Applications, Springer-Verlag; 2003*, p. 457-470.

[11] Cowling PI, Rezig W. Integration of continuous caster and hot strip mill planning for steel production. *Journal of Scheduling 2000*; 3(4):185-208.

[12] Cowling PI, Ouelhadj D, Petrovic S. Multi-agent systems for dynamic scheduling. *Proceedings of the Nineteenth Workshop of Planning and Scheduling of the UK, PLANSIG 2000*; 2000, p. 45-54

[13] Cowling PI, Ouelhadj D, Petrovic S. A Multi-agent architecture for dynamic scheduling of steel hot rolling *Proceedings of the Third International ICSC World Manufacturing Congress, Rochester, NY, USA;* 2001, p. 104-111.

[14] Cowling PI, Johansson M. Using real time information for effective dynamic scheduling. *European Journal of Operational Research 2002;* 139 (2):230-244.

[15] Cowling PI, Ouelhadj D, Petrovic S. A Multi-agent architecture for dynamic scheduling of steel hot rolling, *Journal of Intelligent Manufacturing 2003*; 14:457- 470

[16] Cowling PI, Ouelhadj D, Petrovic S. Dynamic scheduling of steel casting and milling using multi-agents. *Production Planning and Control, Special Issue on the Application of Multi Agent Systems to Production Planning and Control 2004;* 1:1- 11

[17] Soumen Chakrabarti, Martin Van den Berg, Byron Dom "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery" *The Eighth International World Wide Web Conference,* May 11-14, 1999.

[18]. Tang L, Liu J, Rong A, Yang Z. A mathematical programming model for scheduling steel-making-continuous casting production. *European Journal of Operational Research 2000;* 120(1):423-435.